Program Współpraca Polska-RPA	**RAPORT CZĄSTKOWY z realizacji projektu w ramach programu międzynarodowego Polska-RPA**			Narodowe Centrum Badań i Rozwoju
Nr raportu	IR-RATfor5G-02			
Data aktualizacji raportu:	2020.01.01		**Wersja**	2
Numer umowy	PL-RPA2/02/RATfor5G+/2019		**Akronim**	RATfor5G+
Okres realizacji projektu	od	2019.01.01	do	2022.06.30
Tytuł projektu	Technologie dostępu radiowego dla standardu 5G i przyszłych generacji sieci bezprzewodowych			
Tytuł raportu	NOMA/MIMO-NOMA/mmWaves w sieciach IoT			

Artykuł opublikowany w Zeszyty Naukowe UTP Nr 271, Telekomunikacja i Elektronika 24 (2020) str. 45-65
Link: https://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-98ccdbdd-1a65-4c68-a788-335c086ba981

1. Introduction

In general, papers dealing with wireless communications, and in particular the review papers, use many different abbreviations, and this paper is not an exception to that rule. Therefore, to facilitate reading, a list of abbreviations used commonly throughout the paper is given as an appendix. The paper presents an overview of NOMA as an important candidate for multiple-access scheme for beyond 5G. Our contribution addresses NOMA in connection with IoT/mMTC traffic expected to grow largely in the coming years. Moreover, NOMA is a promising technique to help address aspects of task offloading (with MEC/edge computing), incorporate some physical layer security, to mention just a few benefits of using it. The main aim of this paper is to go beyond state of the art available nowadays and indicate directions for NOMA that make it even more attractive in combination with edge computing, cloud-RAN and especially AI/ML to improve future 6G networks. There are various types of NOMA (as already studied by 3GPP in technical reports), and to date, there is no consensus which particular type of NOMA should be supported by next releases of standards (Rel.17 and beyond). However, existing evaluations show that the different techniques do not differ much concerning their validation in the system or link level simulations. However, the authors of this document foresee valuable usage of NOMA together with novel paradigms for mobile systems including for example a cell-free network. Additionally, we believe that concepts like cloud/edge, big data, and virtualization are of the same importance for both domains i.e. IoT platforms as well as the NOMA. Both topics studied in this paper (so IoT in combination with NOMA access technologies) have strong and direct influence on the capacities of NOMA assisted IoT/mMTC future use-cases. From the technical perspective the important

NOMA enabler "[…] *superposition coding allows to exploit the high channel disparities among users, for enhancing the weak users' rates without incurring much degradation of strong users' rates"* [1].

As authors further indicate in [1], integrating NOMA in future systems strongly depend on the capability of the network to address issues related to efficient mitigation and handling of inter-cell interference, as NOMA is very vulnerable to inter-cell interference. Therefore, it should become a strong indication of future direction for NOMA and also cellular networks in general that coordinated (centralized) and virtualized open-RAN networks can be valuable deployment solution for NOMA already providing means for system-level interference removal (i.e. the coordination addressing resource allocation at some central point). The cell-free paradigm has also recently been proposed to overcome limitations of existing cellular systems.

In this paper, authors assume that IoT networks in the future will be augmented with NOMA-capable sink-nodes (proxy or gateway) that can deliver aggregated data from multiple sensors for delivery to the cloud or other server. The sensors themselves will be communicating with each other using various techniques (LoRa, SigFox, Zigbee, etc). NOMA is considered by 3GPP standardization as strong candidate for 6G multiple access technique.

1.1. IoT use-cases

The NOMA in uplink seem to be crucial technique for the IoT networks. To better align NOMA and IoT we present here some high-level view of the existing IoT directions. The report from AIOTI alliance [2] presents various envisioned soon use-cases for IoT together with the corresponding requirements for the network side of 5G. The use-cases considered as most interesting from the perspective of this paper are: smart mobility (e.g. urban driving, vehicle monitoring, car sharing), smart energy, or smart manufacturing. Among emerging topics for IoT are: (i) tactile internet, (ii) ETSI ITS G5 vs LTE-V2X, (iii) 5G non-public networks and network slicing and (iv) 5G in energy industry. In general, the foreseen use-cases can be covered by the existing 5G services (mMTC, URLLC) but there are some cases where existing capabilities foreseen by the IoT use-cases cannot be met with the current 5G features.

The NGIoT Consortium has prepared a review paper [3] where it indicates the seven application domains for IoT, twelve key challenges from economic and policy perspective as well as technological fields where the advances influence the IoT success including: edge computing, 5G, AI and analytics, augmented reality, tactile internet, digital twin and distributed ledgers. All in all, the NGIoT paper recommends to the Horizon Europe program the inclusion of (among others): research to ensure the development of reliable, low-cost, sustainable and scalable IoT networks, leverage the advancements in AI and Ledgers and other technologies to evolve IoT platforms beyond today's limitations as well as develop security-by-design and develop IoT miniaturization, energy harvesting and pervasiveness. In addition, large scale pilots are recommended for the IoT widespread testing and further deployment.

On the other hand, the AIOTI HLA [4] describes the high-level architecture for deploying IoT trials. These documents build on top of the ISO/IEC 42100 standard which captures terms and concepts for IoT architectural models (domain model, functional model, communication model, information model, physical entity model, and integrity model). The main deployment technologies and concepts for IoT HLA deployment are cloud and edge (including OpenFog), big data and virtualization. It is important to notice that **HLA is designed to be a largely distributed system**. Fog approach extends pure cloud into more transversal environment where multiple edges can cooperate horizontally. The general AIOTI functional model however builds on three layers: Application layer, IoT Layer, and Network Layer. The IoT Layer groups IoT specific functions (data storage, sharing) exposed to application layer. While the Network Layer delivers the data plane, and control plane functions. It is important to notice that some devices (constrained) can contain only Application Layer, and Network Layer, whereas the IoT gateway would play the role of “IoT entity”, while other devices may include all the layers in one place.

1.2. NOMA and IoT in standards

It is essential to better understand the standardization perspective behind NOMA and IoT, and especially what are the main overlaps (common points) in the future planning.

The release 15 of the 5G NR, has specified grant-free (GF) transmission in NR to reduce signaling overhead and latency, which is suitable for both URLLC and mMTC, especially in the uplink. In the next 5G NR release standards (Rel.16, Rel.17) in addition to connectivity through the cellular infrastructure, side link connectivity with another IoT device or smartphone will be introduced. Moreover, GF transmission will be extended to support side link transmissions and enhanced with NOMA. Finally, Narrow Band IoT (NB-IoT) and LTE-M will be integrated with 5G NR to provide dedicated MTC services [5]. The summary of Rel.16 work items mentions enhancements to mMTC in the recently published specification [6], these enhancements among all include enhanced coverage RAN feature, control of user data rate sent to/from UE, control plane congestion control, inter-UE QoS for NB-IoT, etc.

There are also “5G lite” solutions for Rel.16/17 that may be an interesting point in the delivery of NOMA as they are targeting some “balance” between 5G use-case and deliver solution that resembles LTE from performance perspective but already uses the 5G signaling and architecture, which basically allows for complementing or even replacing the NB-IoT and LTE-M in the future.

A NOMA background especially for the uplink direction together with comprehensive simulation results of NOMA (including link and system levels) has been presented in the 3GPP document [42]. The generic NOMA transmitter side processing diagram as well as receiver are presented there. Receiver complexity has also been analyzed and compared for different receiver types. Moreover 35 different test scenarios are provided with detailed settings for LL simulation, including: carrier frequency, 5G use-case, SNR distribution, waveform, channel model, TBS size, and #UEs. In general, results show more differences between NOMA schemes for larger TBS sizes, as for smaller TBS performance differences are small for all NOMA schemes. Performance degradation is also identified for realistic channel (in the range of 2-5dB). In a system level simulation all three use-cases are considered (together with detailed parameters): eMBB, URLLC, and mMTC. Similar analyses for multiple-access technologies including NOMA are also included in LL/SLS simulation results provided in earlier report for Release 14 [41]. The green elements in the figure depict modifications required by NOMA in the physical layer transmitter side.

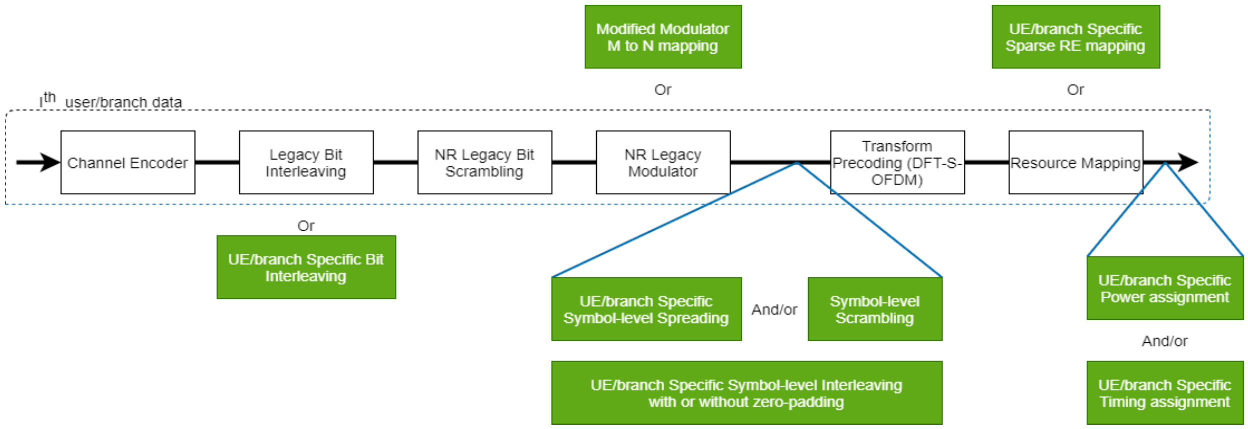

Figure 1Generic NOMA transmitter (source [42])

Besides that concerning the IoT terminals the document [43] captures various features of mMTC terminals and their modifications to reduce cost and improve coverage along with various hardware simplifications that will enable production of low-cost MTC user equipment (UE). The next section introduces some background on the IoT traffic characteristics and modeling that might be required for simulations with NOMA.

1.3. IoT traffic characteristics

The comprehensive survey in [7] presents an overview of the most significant 5G usage scenarios and traffic generation models. These environments and traffic models will allow 5G stakeholders and researchers to evaluate the performance of 5G solutions under the most critical requirements. The requirements related to MTC have been cited in table **Błąd! Nie można odnaleźć źródła odwołania.**.

Table 1-1 MTC characteristics (after [7])

Parameter	Statistical characterization
	Smart metering
Description	Periodic or on-demand metering reports
Session density	<50 times/day/metering device
Bytes/session	Tens of bytes
Deployment	Urban/suburban
Mobility	Stationary
Machines/km²	<10000 (urban), <1000 (suburban)

	Telemedicine
Description	Periodical patient monitoring; also urgent calls or some control data
Session density	<500 times/day
Bytes/session	Tens of hundreds of KBs
Deployment	Urban/suburban
Mobility	Stationary/mobile
Machines/km²	<1000

	Home automation
Description	Home sensing and control, home security
Session density	<100 times/day/home gateway
Bytes/session	Hundreds of bytes
Deployment	Urban/suburban
Mobility	Stationary
Machines/km²	<10000 (urban), <1000 (suburban)

mMTC	**Fantastic-5G** Sensor networks: Constant packet generation intervals (uplink) [57]
	Metis-II Massive distribution of sensors and actuators: Bursty traffic FTP model 3, file size 125 bytes, IAT down to 1s
	3GPP Massive connection: Non-full buffer with small packets

Family	Category	Use case	User requirements	System requirements
Massive Internet of Things	Massive low-cost/long-range/low-power MTC	Smart wearables (clothes) Sensor networks	*Data rate*: low (typically 1-100kbps) *E2E latency*: seconds to hours	*Connections*: 200 000/km² *Traffic*: not critical
	Broadband MTC	Mobile video surveillance	Same as *Broadband access in dense areas and 50+Mbps everywhere* categories	

The survey in [7] considers the requirements for use-cases, provides deployment scenario characterization and their traffic models. These topics are studied across multiple organizations (3GPP, IEEE, ITU) as well as industry associations (5G-PPP, NGMN, TIA). In the 4G networks at cell level, it is expected that each household in a cell may have up to 40 MTC devices and the household density per cell is according to the assumptions in Annex A of [43]. The resulting MTC device density per cell is provided as well, and the packet size is foreseen to be ca. 20-40 bytes. Also the size of bandwidth (BW) is limited. The industrial use-cases are described in [8]. Example parameters characterizing URLLC traffic in 5G are presented in the Table 1-2.

Table 1-2 URLLC traffic characteristics

Use case (high level)		Availability (%)	Cycle time (ms)	Typical payload size	# of devices	Typical service area
Motion control	Printing machine	>99.9999	<2	20 bytes	>100	100 m x 100 m x 30 m

	Machine tool	>99.9999	<0.5	50 bytes	~20	15 m x 15 m x 3m
	Packaging machine	>99.9999	<1	40 bytes	~50	10 m x 5 m x 3 m
Mobile robots	Cooperative motion control	>99.9999	1	40-250 bytes	100	< 1 km^2
	Video operated remote control	>99.9999	10-100	15-150000 bytes	100	< 1 km^2
Mobile control panels with safety functions	Assembly robots or milling machines	>99.9999	4-8	40-250 bytes	4	10 m x 10 m
	Mobile cranes	>99.9999	12	40-250 byes	2	40 m x 60 m
Process automation (process monitoring		>99.99	>50	Varies	10000 devices per km^2	

The graphical summary of key characteristics of mMTC traffic is presented in the Figure 2.

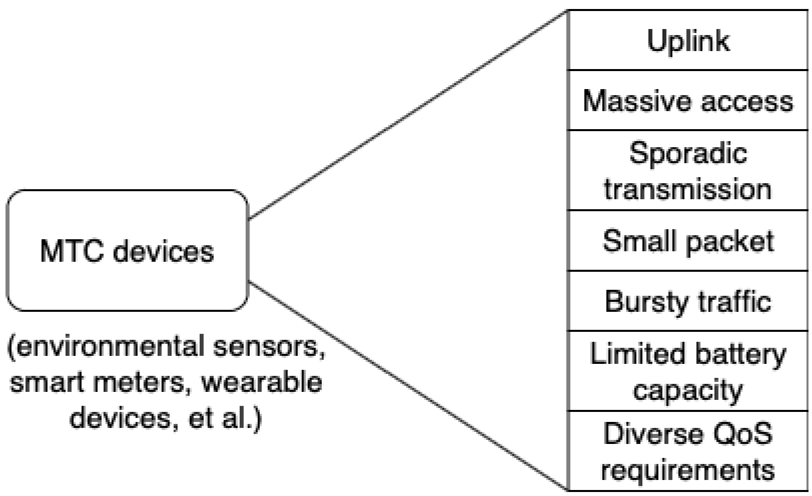

Figure 2 mMTC characteristics (source [23])

In the 5G triangle URLLC and mMTC are defined through QoS requirements. In [9], authors analyze the profiles and traffic specifications of mMTC and URLLC. Authors present comparison between the two service types (see Table 1-3) but also provide overview evolution of enhancements for mMTC in new releases of 3GPP specifications (Rel.15 and 16). The C-RAN coexistence with the service based architecture of the core network, slicing, orchestration and automation are listed as important features to support future IoT networks.

Table 1-3 Characteristics of mMTC/URLLC

Specifications	**mMTC**	**URLLC**
Connection Density	Up to 10^6/km^2	Comparably low
Power Consumption	Extreme low	Insensitive
End to End Latency	Insensitive	$1-10$ ms
Reliability	Typical BLER 10^{-1}	Up to BLER 10^{-9}
Payload Size	Small	Small to large
Bandwidth	Narrowband	Wideband
Numerology	3.75 KHz 15KHz	$2^n \times 15$KHz

Moreover according to the authors, the URLLC-mMTC co-existence can also be considered by employing power-domain NOMA within a shared resource block, where each sub-carrier can be shared by delay sensitive and delay tolerant devices. In [10], authors highlight an mMTC traffic property that a set of devices transmitting in each radio resource is random and unknown. Moreover, as mMTC has a constant rate the authors focus on an arrival rate of mMTC devices that can be supported as the main performance criterion. It is mentioned that the SIC decoding improves mMTC performance as it can leverage power imbalances to sequentially improve the reliability of simultaneous mMTC transmissions. Practical considerations for suitability of using SIC for systems where both eMBB and URLLC traffic types are processed in parallel are also studied in [50]. Technical report by NTT Docomo in [11] shows field test results that proves that latency and reliability requirements of URLLC can be met in 5G network.

1.4. IoT architectures beyond 5G

Software-based architectures of 5G and beyond bring increased energy consumption due to COTS servers used in place of domain specific chips. That is why 6G networks require new computing paradigm to support benefits of software defined networks without bearing the costs of energy consumption [8]. IoT virtualization according to ETSI 103 527 [12] requires “that the elements of an IoT system <u>can work in a fully interoperable, secure and dynamically configurable</u> manner with other elements (devices, gateways, storage, etc.) that are deployed in different operational and contractual conditions”. Considering the IoT virtualization proposed by AIOTI alliance the microservices-based functional architecture of IoT entity (that can be composed of multiple IoT devices) is presented in [4]. The two vertical functions that are related to cross-layer are security and management. Security is a global requirement for every IoT system that impacts every layer of the system architecture.

It is highlighted in [4] that microservices approach is the direction to be considered to enable IoT virtualization. Microservices allow for fine-grained decomposition of processing required in the IoT domain to get most benefits from virtualized processing of data. The architecture presented in Figure 3 becomes main component of the “IoT Layer” inside the AIOTI HLA architecture.

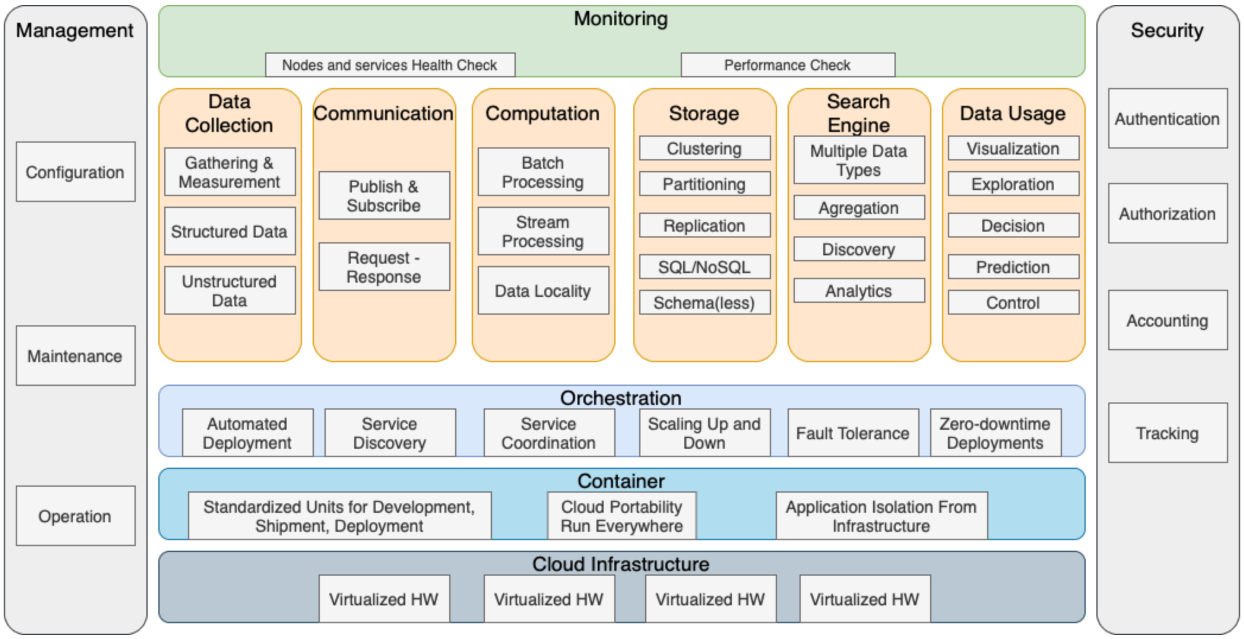

Figure 3 High level architecture (source [4])

The AIOTI also defines a high-level architecture of virtual objects based on the iCore project [13].

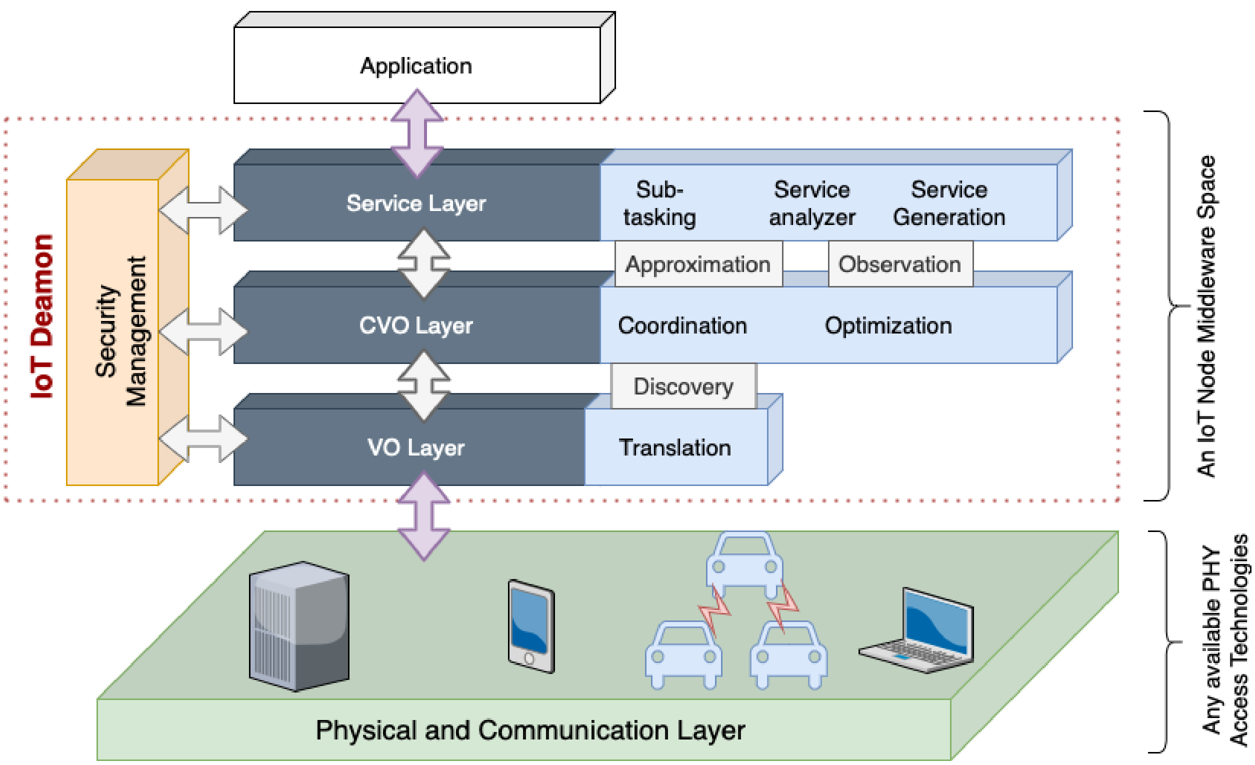

Figure 4 Service based abstraction of IoT device (source [13])

As indicated in [4] "*[…] Virtual Objects and Virtual Composite Objects are a method to introduce an abstraction layer through which the devices and groups of devices present themselves to the network. Instead of a collection of very small and specific functionalities, the devices are grouped together to form complete virtual devices*". Such architecture is presented in Figure 4 after [4] and

[13]. Another reference model for IoT can be referenced from the ITU-T Y.4000. The ITU-T Y.4000 model has been mapped to the AIOTI HLA architecture. Another approach is the oneM2M common services entities (CSE) which provide IoT functions to oneM2M AEs (application entities) via APIs. The purpose and goal of the oneM2M partnership is to develop technical specifications which address the need for a common M2M Service Layer that can be embedded into various hardware and software, and relied upon to connect the various types of devices in the field with M2M application servers. In the view of oneM2M there is common API that offers IoT functionalities to applications. The service layer is dependent on the network services entity (NSE) layer that provides services like: location, device triggering, sleep modes etc.
The main reference points are: Mcc (between CSE), Mca (between AE and CSE), Mcn (between CSE and NSE). For the Mcc and Mca, three protocols are defined: CoAP, MQTT, Websockets and HTTP. However, the Mcn reference point (towards network) is specified by 3GPP. The three-tier architecture is defined by the Industrial Internet Reference Architecture (IIRA) which is based on an open architecture. The mapping of this architecture on the IIRA is presented in Figure 5.

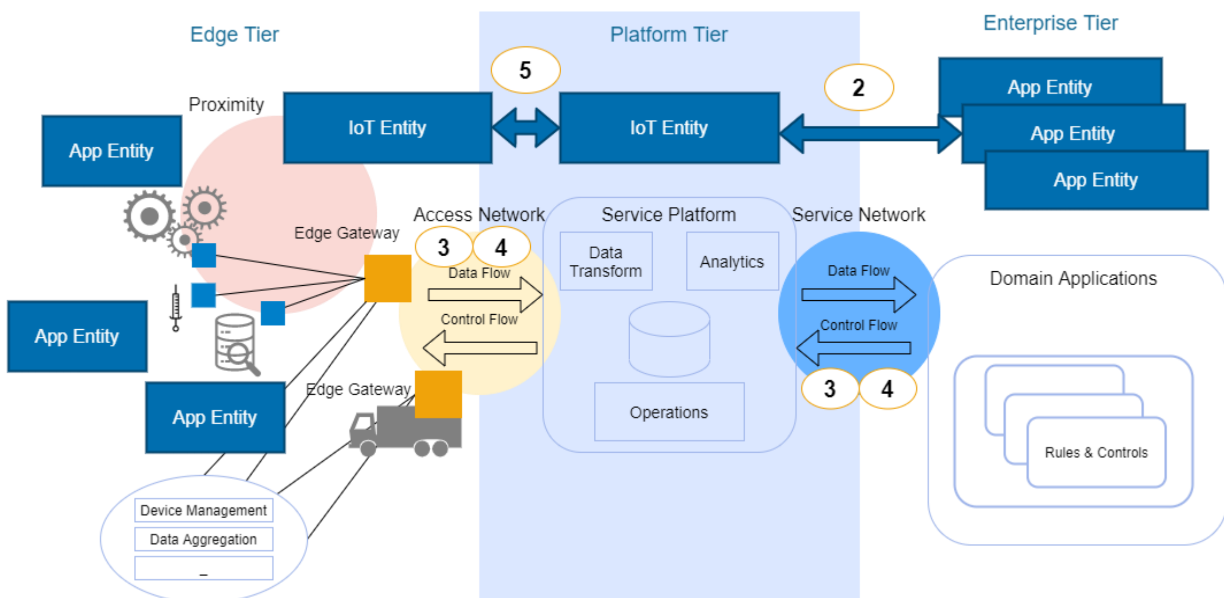

Figure 5 Open architecture of IoT by the IIRA (source [4])

The edge gateways are mapped into "IoT Entities" in this architecture. Interaction with the network is performed using the interfaces "3" and "4" from the AIOTI HLA. It is worth noting that in the architecture above the IoT entity can be IoT gateway or the nodes themselves. It is important to highlight that, due to significant number of IoT related communication technologies (standard based like: LoRa, Sigfox, NB-IoT, LTE-M, Zigbee as well as proprietary: Abax2 etc), there is an important need to bridge between such specialized technologies. In this paper, we highlight that NOMA can be useful at various levels of communication hierarchy. Especially, it is essential to notice that capabilities of end IoT devices can be varying. Hence, using gateways may be often the right solution to follow when aggregating and delivering to mobile networks (and then clouds for application traffic). IoT gateways as explained in the architectural figures from AIOTI, ITU or ETSI can be the places where NOMA could be deployed in both uplink link to the mobile network as well as downlink link towards mobile terminals under IoT standard.

2. NOMA/IoT in C-RAN networks

Cloud-RAN, centralized-RAN, and virtual-RAN are different deployment options that appear on the horizon as a consequence of standards supporting virtualization, as well as prevailing trend in moving applications and services into a cloud. Different workloads in the cloud span from user-facing services (like computation services for IoT, various utility services, etc.), security services (like processing of DPI, signal processing for attack identification and prevention), as well as infrastructure services (e.g. virtual RAN, 5G RAN). There is essential difference between the various RAN deployment options regarding requirements for QoS and computation resources. The virtual 5G RAN deployment already challenges existing cloud data centers as it requires special support, like access to acceleration technologies (GPU, FPGA, Smart NIC), real-time processing for the 3GPP protocol layers like e.g. PHY (especially Low-PHY considering the functional splits of 3GPP [14]). According to [15] the "*[...] exploiting interference that affects UL users can significantly improve their QoS and spectral efficiency. In C-RAN, multi-cell NOMA allows such interference exploitation.*" The Figure 6 provides conceptual view on C-RAN based, multi-cell (or cell free) NOMA.

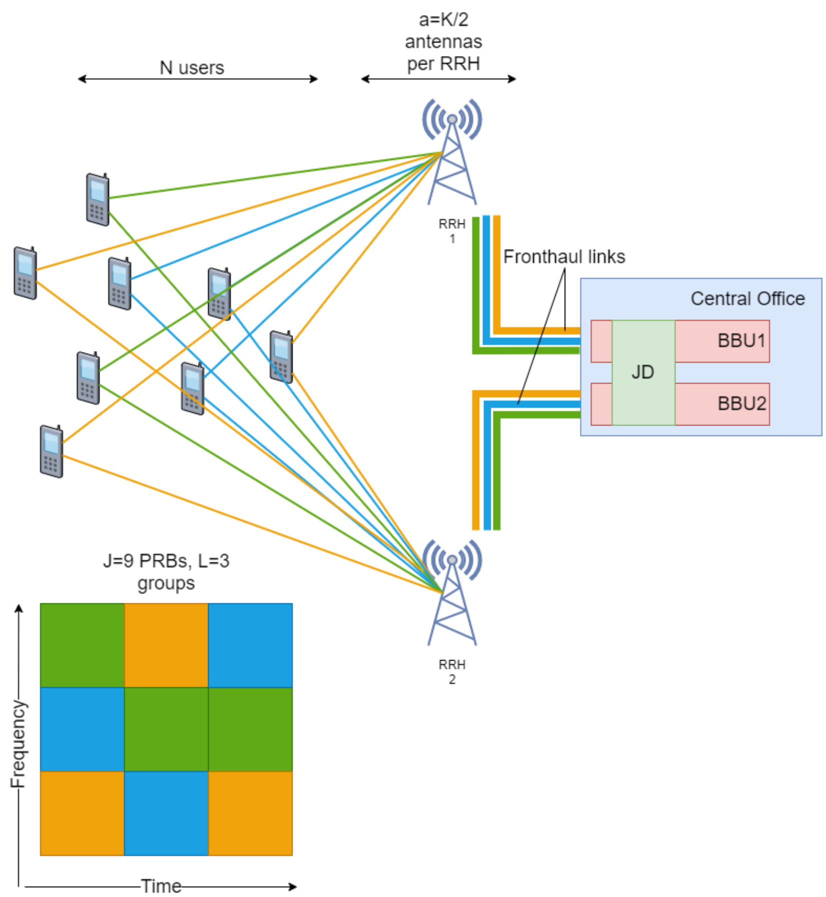

Figure 6 Uplink NOMA in C-RAN conceptual diagram (source [15])

The main standard body influencing virtualization based developments is ETSI with its NFV group of specifications. The NFV architecture allows unifying the view on "multitude of workloads" and classifies it into: virtual functions (VNF) and physical functions (PNF). The first group refers to any workload that can be turned into virtual machines or containers. By doing so, they become HW independent (NFVI is performing the role of a runtime environment in this scenario) and thus multitude of general-purpose processors can be handling such a workload. The functions that cannot be moved, or are provided in a more legacy packaging (e.g. embedded HW) are also considered by the NFV specifications for completeness. However not all of the 3GPP radio stack functions (whether VNF or PNF) can meet its performance targets without support of accelerators of various kinds in order to meet stringent demands for the processing.

2.1.1. IoT Virtualization

IoT virtualization trend goes along the sophisticated capabilities provided by the cloud computing. **IoT systems require high degree of availability, adaptability, and flexibility** – and cloud models have been designed to serve such requirements. Benefits of virtualization in the context of IoT are indicated further in [4]: (i) rapid service innovation through service-based deployment and operation of IoT devices, (ii) improved operational efficiencies resulting from common automation and operating procedures, (iii) reduced power usage by migrating workloads and powering down unused HW, (iv) greater flexibility on assigning IoT virtualized functions and objects to HW, (v) improved capital efficiencies compared to dedicated HW implementation.

An essential research and development work of the recently finalized H2020 project Coral [16] was the software virtualization of the IoT communication stacks of multiple RATs. In the course of the project the following radio access technologies (RAT) were virtualized:

- IEEE 802.15.4: full-stack implementation supporting 3 frequency channels from PHY layer to application layer, with bi-directional communications between the softwarized communication stack function in the Edge and commercial IoT devices (i.e. Zolertia firefly [17]).
- LoRa: PHY and MAC layer implementation with bi-directional communications between the virtualized communication stack function in the Edge and commercial IoT devices (i.e. Pycom FiPy [18])
- NB-IoT: downlink PHY (NPSS, NPSCH) implementation with simplified upper layer implementation which supports sending signals and messages from the virtualized communication stack function in the Edge and a self-developed SDR-based NB-IoT receiver.

The physical layer processing is done at the Edge and the radio head (RRH) is responsible for the configuration and management of the Software Defined Radio (SDR). The SDR converts radio channel information into digital streams of In-Phase and In-Quadrature samples. These samples need to be transported to and from the Edge for the receive and transmit data flow chains respectively. The protocol selected for this transmission was IP and the throughput levels of e.g. 802.15.4 was 128Mbps. The figure below presents the block diagram of the proof-of-concept implementation in of the “MultiRAT” use-case in the project.

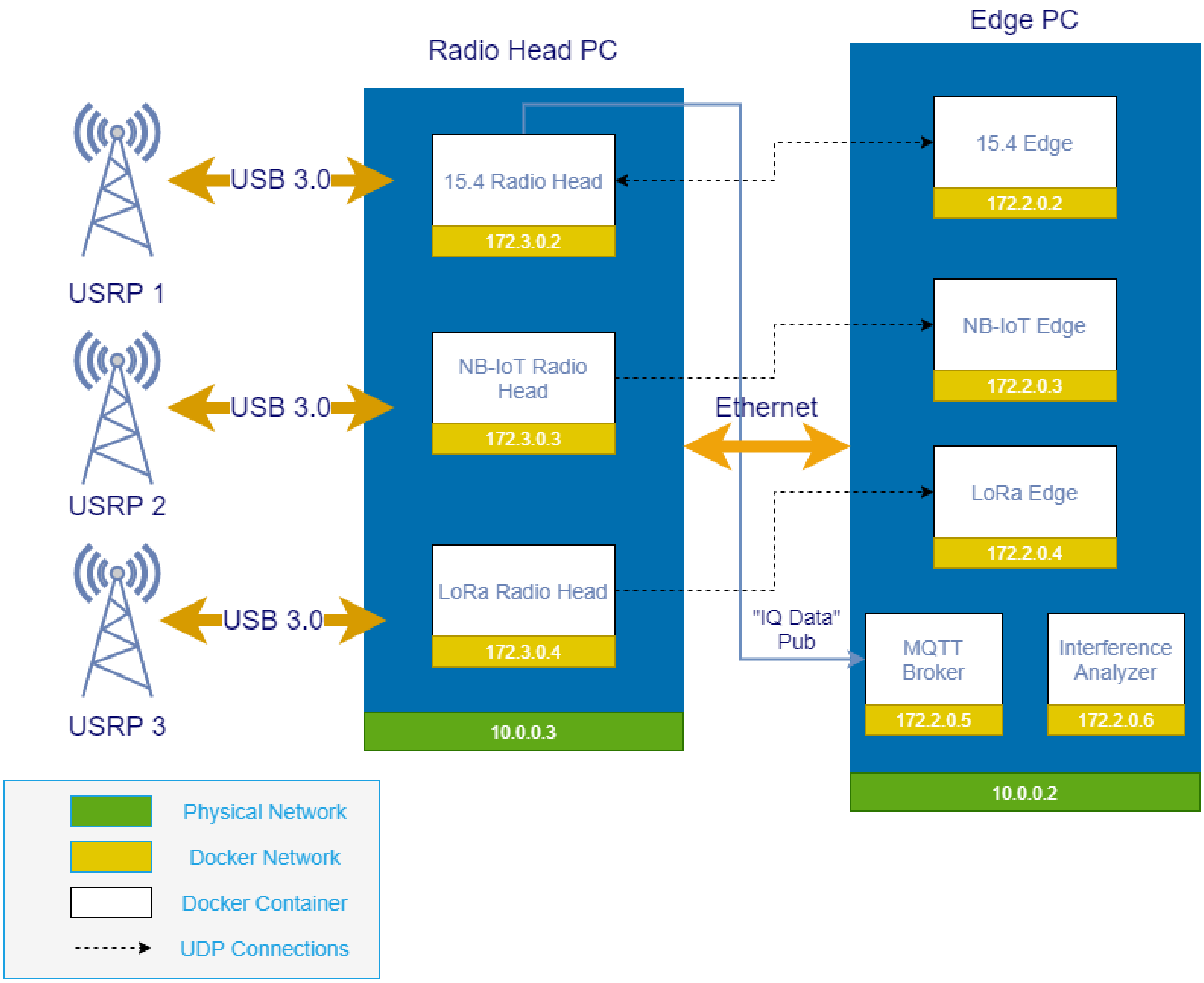

Figure 7 Coral project IoT testbed (Source [19])

Detailed description of this testbed is presented in the D4.2 deliverable of the 5G Coral [19], where the HW and SW components have been provided. Some essential IoT virtualization use cases are mentioned by ETSI in [12]: (a) horizontal up and down auto-scaling of cloud resources in response to time-varying amount of data transmitted by IoT devices, (b) no single point of failure – to prevent from failures when server or gateway goes down and (c) data privacy – related to data anonymization and exchange of data between various devices. These use-cases highlight the crucial enablers behind virtualization of IoT architectures.

From the perspective of resource management in virtualized IoT systems, the following functional aspects need to be considered: (i) multi-tenancy, (ii) massive data processing. Similarly, the non-functional requirements need to be considered including: (i) high-throughput, (ii) high-availability, (iii) low-latency, (iv) edge-computing, and (v) security. Whereas high-throughput refers to how many jobs can be completed over a long period of time instead of how-fast. The high-availability translates to the fact that service should be designed so that the process can be restarted at any time with no data loss (e.g. with the use of multiple service instances managed by load-balancer). **The low-latency means that routing of IoT application messages may be happening in (near) real-time or reducing network latency by moving computing resources closer with**

the use of Edge Computing. Additionally, the MapReduce concept is introduced which shifts the processing of data closer to the data origin – thus only the results are returned and not much traffic is generated. Moreover, the edge functionality can reside on either (a) device edge – where the sensors talk to a local edge (IoT Gateway) device which manages connectivity with the cloud or (b) cloud edge – where the processing of data happens inside edge nodes distributed by a cloud provider closer to the user.

Another natural approach to virtualization is to follow the ETSI NFV ISG architecture. It is the same architecture that is proposed by O-RAN and 3GPP in approaching the cellular-networks virtualization based on open-protocols. The benefit of following this architecture is the consistency of reference points between cellular and IoT communication paradigms.

Specific challenge for IoT is the coexistence of multiple verticals (or application-domains) and IoT flavours (e.g. LoRA, LTE-M, NB-IoT) on top of existing network infrastructure (E2E). This approach is motivated by the need to isolate IoT data, data processing, and delivery across multiple operators, by assuring both security and QoS requirements. It has been proven already in [20] that where heterogeneous networks and differentiated services are concerned, building dedicated networks for each service type, can be the most efficient solution. In such a case, the "slicing concept" becomes useful. Slicing as defined by NGMN [21], 3GPP [22] "[..] enables operators to create networks customized to provide optimized solutions for different market scenarios demanding diverse requirements, e.g. in the areas of functionality, performance, and isolation. This is a key requirement from HLA and related IoT use cases, and stakeholders such automotive, energy, cities, etc."

It is interesting to note that the IoT transmission shares the same spectrum, and the same infrastructure as the cellular network, and adding IoT transmission can even help the cellular transmission due to introduction of multipath diversity [23]. The backscatter communications back-up the symbiotic radio (SR) concept, which supports massive access from IoT devices by passively reflecting the signals received from the cellular TXs such as BS or mobile stations (MS).The use of NOMA in combination with fog-RAN (FRAN) is studied in [1], where FRAN is described as an alternative to the C-RAN deployments. The main difference between FRAN and CRAN is reduction of requirements for fronthaul in FRAN as processing is shifted to the edge. The drawbacks of cloud processing where baseband units (BBUs) shared in the cloud manage interference and resource allocation is the capacity limited fronthaul links, that may introduce transport delays. To overcome CRAN limitations FRAN is partially moving network intelligence, i.e., cloud computing and storage capabilities, closer to the network edge. The access points in FRAN (Fog Access Points - FAPs) can perform distributed signal processing and radio resource allocation. There are references that show utilizing FRAN/FAPs where RRM and interference mitigation can be performed with the use of NOMA [24][25][26]. Exploiting NOMA in downlink transmission of FAPs increases user fairness without sacrificing data rates. Additionally, combination of D2D and NOMA under FRAN is studied in literature [27][28] for jointly providing high data rates and low latencies in eMBB. In the mMTC scenario NOMA can boost uplink transmission by increasing a number of devices (per resource block - RB) that can be connected to the network in grant-free access. "[..] *In the downlink, NOMA multiplexes multiple user messages on the same basic RB unit. In the uplink, NOMA allows several users to simultaneously access the same basic RB unit without collisions. Moreover, cloud-based optimized packet* *scheduling, radio resource allocation (RRA) and interference mitigation (IM) can jointly enhance the network reliability metrics of the users* [1]".

Unlike the conventional cloud computing operated in the remote cloud that suffers severe transmission latency via the Internet, MEC offers cloud computing capabilities at the edge of radio access network (e.g., at small-cell BSs) in close proximity to NB-IoT devices [29]. Through bringing intensive computation tasks from NB-IoT devices to MEC units, the low-latency as well as reliable computing services can be implemented for NB-IoT devices.

2.1.2. Resource management and 5G use-cases

Grant-free NOMA is a generic technology that can bring benefits to mMTC, URLLC, eMBB small packet and two-step random-access channel scenarios [30]. However, if two or more users select the same resource for transmission, a collision occurs. Under this scenario, the receiver is unable to decode the data of users sharing the same RB. There are two ways of performing grant-free access: 1) UE's resources are pre-configured and periodically allocated, and each time when a packet arrives, the UE would choose the nearest allowable time-frequency resource for the uplink transmission, which is called semi-persistent-scheduling (SPS) based grant-free; 2) UE can randomly select a resource at any time for uplink transmission, leading to contention-based transmission.

URLLC like requirements are also considered in [31] where authors deal with power control for delay-bounded IoT applications. Typical emerging IoT applications require a latency from 0.25 ms to 10 ms and an outage probability (or packet loss rate) in the order of 10E–03 to 10E–09 [32]. Often IoT devices like UAV (or other gateway nodes) simultaneously provide service to many battery powered devices, with limited bandwidth resources under statistical delay QoS.

Short packet transmission in absence of a closed-loop control of GF-NOMA is addressed by deep learning in [33]. A remedy here is the open-loop selection of transmit power from the pool by users solely based on their communication distance. Each IoT user acts as an agent and learns policy by interacting with wireless environment. To prevent Q-learning overestimation problem a double DQN based GF-NOMA is proposed. It converges faster than Q-learning under changing environments due to limiting action space based on previous learning. Information about channel can be calculated via IoT users' geographical information and practical statistical models without information exchanges, which enables an open-loop control. Authors in [34] designed users and sub-channel clusters in a region, where number of users compete in a GF manner for several available sub-channels in each region. The formulated long-term cluster throughput problem is solved via DRL based GF-NOMA algorithm for optimal sub-channel and power allocation.

Authors in [33] "[…] propose a multi-agent deep Q network (DQN) and double DQN based GF-NOMA algorithm for prototype power pool design, where the BS broadcasts this pool to all IoT users to avoid acquiring CSI. Each IoT user can randomly select one power level for transmission that reduces complexity at BS and avoid massive information exchange between IoT user and the BS. Power selection from this well-designed prototype power pool guarantees distinct received power levels at the BS for successful SIC processes and reduces collision probabilities by allowing pilot sequence reusing […]. Authors consider uplink transmission in IoT networks with the traffic model of packets following the Poisson distribution. Further, authors divide the cell area into different layers and design a layer-based transmit power pool prototype via multi-agent reinforcement learning (MARL). In the proposed framework, data transmitting IoT users select a transmit power based on their communication distance (layer) from the well-designed prototype power pool for GF-NOMA transmission, without any information exchange between IoT user and the BS." IoT user acts as a learning agent and interacts with the environment. After learning from its mistakes, the IoT users in each layer find out the optimal transmit power level that maximizes network throughput.

In the downlink, the SIC decoding order is fixed. The user with the stronger channel must first decode the message intended for the weaker user, subtract the corresponding signal, and can then decode its own message. The weaker user will only be able to decode its own message. Contrary to that, the decoding order in uplink NOMA can be chosen arbitrarily. Many works related to uplink NOMA [45][46][47][48][49] suggest that the signal from the stronger user should be decoded first, such that the weaker user's signal is interference-free.

In [39] authors investigate the achievable link-layer rate of a two-user NOMA with short-packet communications i.e. the downlink URLLC case, where the Shannon capacity brings too loose bound. Specifically, they formulate the effective capacity of the strong and weak users under heterogeneous delay QoS requirements. The overall reliability, which is the combination of the transmission error probability and the queueing delay violation probability, is investigated. Authors also derive closed-form expressions for the individual effective capacity of the two NOMA users. The important is the calculation of an achievable data rate:

$$r_i = \ln(1+\gamma_i) - \sqrt{\frac{\delta_i}{n}}\,Q^{-1}(\varepsilon_i)$$

as well as the queueing model of the NOMA communication for two users. Based on this assumption effective capacity of both: weak and strong users are derived.

Whereas in [44] authors study the effect of delay in uplink under imperfect CSI and finite-length coding. They indicate that delay violation as important metric in URLLC/mMTC scenarios will be dependent on the SIC decoding order. In order to determine the delay performance of NOMA systems in the presence of decoding errors and error propagation, one must first analyze the decoding error probabilities due to imperfect CSI and finite block length channel coding. Authors conclude that even under realistic assumptions, NOMA may be suitable for low-latency communications, but only when joint decoding is used and only when there is a large difference between the two users' average SNR values. However, joint decoding may be difficult to implement in practice, especially for low-latency systems. With SIC decoding, NOMA often performs worse than OMA when considering low-latency communications with more realistic system effects.

The authors in [40] have applied interference cancellation schemes and superposition coding at the NOMA receiver, which can help with multiplexed multiple users on the same subchannel. There are some iterative algorithms with - according to the authors - "guaranteed convergence to deliver a competitive suboptimal solution". The performance evaluations presented there indicate that the effectiveness of the proposed algorithm is better than other resource allocation schemes in NOMA or OFDMA system.

3. Conclusions

This paper has provided an overview of selected IoT architectures, relevant requirements towards 5G (and beyond) systems as well as indication of NOMA techniques suitability as a multiple access technique and recent developments in the area of radio interface for IoT coexisting with 5G (and beyond) mobile networks . The motivation behind this paper was also a desire to further understanding of cross-layer design issues for NOMA and IoT, which has been used in shaping the contents and structure of this paper. An important factor is that NOMA is considered as strong candidate for beyond-5G standards in 3GPP although decision has not been made to date. Hence, NOMA seems to be in a phase of intense research, especially when considering topics important in connection with 6G. Developments related to IoT are now expanding at a large pace, and NOMA is an important enabler that can influence efficiency of all the three types of 5G/6G services. Although the mMTC and URLLC are most important considering number of expected mobile devices that will be growing with IoT popularity. It should be stressed here that information provided in the paper is valid as of the date of its submission and given a rapid pace of research and development in the areas of the IoT and the radio interface for next generations of mobile networks the paper might soon has just a historical value.

4. References

[1] M. Kaneko, I. Randrianantenaina, H. Dahrouj, H. Elsawy and M. -S. Alouini, "On the Opportunities and Challenges of NOMA-Based Fog Radio Access Networks: An Overview," in IEEE Access, vol. 8, pp. 205467-205476, 2020, doi: 10.1109/ACCESS.2020.3037183.

[2] AIOTI (2019), "IoT Relation and Impact on 5G", [Online], Available: https://aioti.eu/wp-content/uploads/2019/03/AIOTI-IoT-relation-and-impact-on-5G-190308-R2-published.pdf

[3] NGIoT, (2019), "*Building a roadmap for the next generation internet of things. Research, innovation and implementation 2021-2027*" [Whitepaper], https://www.ngiot.eu/wp-content/uploads/sites/26/2019/09/NGIoT_scoping-paper.pdf

[4] AIOTI, (2020), "AIOTI - High Level Architecture. Release 5.0" [Whitepaper] https://aioti.eu/wp-content/uploads/2020/12/AIOTI_HLA_R5_201221_Published.pdf

[5] Mahmood N.H, et al. (2020). White Paper on Critical and Massive Machine Type Communication Towards 6G. ArXiv, arXiv/2004.14146v2.

[6] 3rd Generation Partnership Project (3GPP), 2020, TR 21.916 „Summary of Rel-16 Work Items", Release 16 (v1.0.0)

[7] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz and J. M. Lopez-Soler, "A Survey on 5G Usage Scenarios and Traffic Models," in IEEE Communications Surveys & Tutorials, vol. 22, no. 2, pp. 905-929, Second quarter 2020, doi: 10.1109/COMST.2020.2971781.

[8] You, X., Wang, CX., Huang, J. *et al.* Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts. *Sci. China Inf. Sci.* **64,** 110301 (2021). https://doi.org/10.1007/s11432-020-2955-6

[9] S. R. Pokhrel, J. Ding, J. Park, O. -S. Park and J. Choi, "Towards Enabling Critical mMTC: A Review of URLLC Within mMTC," in IEEE Access, vol. 8, pp. 131796-131813, 2020, doi: 10.1109/ACCESS.2020.3010271.

[10] P. Popovski, K. F. Trillingsgaard, O. Simeone and G. Durisi, "5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View," in IEEE Access, vol. 6, pp. 55765-55779, 2018, doi: 10.1109/ACCESS.2018.2872781.

[11] Okumura, Yoshihisa Kishiyama Yukihiko. "Field Experiments on 5G Ultra-Reliable Low-Latency Communication (URLLC).", NTT Docomo Technical Journal, vol.20, no.1 (2018).

[12] ETSI TR 103 527 (the present document): "Virtualized IoT Architectures with Cloud Back-ends", [Online], Available: https://www.etsi.org/deliver/etsi_tr/103500_103599/103527/01.01.01_60/tr_103527v010101p.pdf

[13] iCore Deliverable 2.5 (D2.5), Final architecture Reference Model, [Online], Available: https://cordis.europa.eu/docs/projects/cnect/8/287708/080/deliverables/001-20141031finalarchitectureAres20143821100.pdf

[14] L. M. P. Larsen, A. Checko and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," in IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 146-172, Firstquarter 2019, doi: 10.1109/COMST.2018.2868805.

[15] Dora Boviz. Communications multi-utilisateurs dans les réseaux d'accès radio centralisés : architecture, coordination et optimisation. Autre. Université Paris-Saclay, 2017. Français. ffNNT : 2017SACLC035ff. fftel-01591285f
[16] "D2.2–Refined design of 5G-CORAL edge and fog computing system and future directions". [Online]. Available: http://5g-coral.eu/?page_id=37
[17] FIREFLY - Zolertia, https://zolertia.io/product/firefly/
[18] Pycom, FiPy, https://pycom.io/product/fipy/
[19] 5G-CORAL Project, "Deliverable D4.2 - 5G-CORAL Proof of Concept and Future Directions", August 2019. [Online]. Avalaible: http://5g-coral.eu/wpcontent/uploads/2019/09/D4.2_FINAL.pdf
[20] M. A. Callejo-Rodriguez et al., "EuQoS: End-To-End QoS over Heterogeneous Networks," 2008 First ITU-T Kaleidoscope Academic Conference - Innovations in NGN: Future Network and Services, Geneva, Switzerland, 2008, pp. 177-184, doi: 10.1109/KINGN.2008.4542264.
[21] NGMN (2021), "5G smart devices supporting network slicing" [Whitepaper]. NGMN, https://www.ngmn.org/wp-content/uploads/201214_NGMN_5G_SmartDevicesSupportingNetworkSlicing.pdf
[22] 3rd Generation Partnership Project (3GPP), 2018, TR28.801, "Study on management and orchestration of network slicing for next generation network", Release 15
[23] Liu W C, Liang Y C, Li Y H, et al. Backscatter multiplicative multiple-access systems: fundamental limits and practical design. IEEE Trans Wirel Commun, 2018, 17: 5713–5728
[24] I. Randrianantenaina, M. Kaneko, H. Dahrouj, H. ElSawy and M. -S. Alouini, "Interference Management in NOMA-Based Fog-Radio Access Networks via Scheduling and Power Allocation," in IEEE Transactions on Communications, vol. 68, no. 8, pp. 5056-5071, Aug. 2020, doi: 10.1109/TCOMM.2020.2988564.
[25] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang and A. Nallanathan, "Resource Allocation in NOMA-Based Fog Radio Access Networks," in IEEE Wireless Communications, vol. 25, no. 3, pp. 110-115, JUNE 2018, doi: 10.1109/MWC.2018.1700326.
[26] H. Zhang, Y. Qiu, X. Chu, K. Long and V. C. M. Leung, "Fog Radio Access Networks: Mobility Management, Interference Mitigation, and Resource Optimization," in IEEE Wireless Communications, vol. 24, no. 6, pp. 120-127, Dec. 2017, doi: 10.1109/MWC.2017.1700007.
[27] M. Kaneko, I. Randrianantenaina, H. Dahrouj, H. Elsawy and M. -S. Alouini, "On the Opportunities and Challenges of NOMA-Based Fog Radio Access Networks: An Overview," in IEEE Access, vol. 8, pp. 205467-205476, 2020, doi: 10.1109/ACCESS.2020.3037183.
[28] Driouech, S.; Sabir, E.; Ghogho, M.; Amhoud, E.-M. D2D Mobile Relaying Meets NOMA—Part I: A Biform Game Analysis. *Sensors* **2021**, *21*, 702. https://doi.org/10.3390/s21030702
[29] L. P. Qian, A. Feng, Y. Huang, Y. Wu, B. Ji and Z. Shi, "Optimal SIC Ordering and Computation Resource Allocation in MEC-Aware NOMA NB-IoT Networks," in IEEE Internet of Things Journal, vol. 6, no. 2, pp. 2806-2816, April 2019, doi: 10.1109/JIOT.2018.2875046.
[30] L. Tian, C. Yan, W. Li, Z. Yuan, W. Cao and Y. Yuan, "On uplink non-orthogonal multiple access for 5g: opportunities and challenges," in China Communications, vol. 14, no. 12, pp. 142-152, December 2017, doi: 10.1109/CC.2017.8246331.
[31] C. Xiao, J. Zeng, B. Liu, X. Su and J. Wang, "Cross-Layer Power Control for Uplink NOMA in IoT Applications with Statistical Delay Constraints," 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 2018, pp. 1-7, doi: 10.1109/GLOCOM.2018.8647452.

[32] P. Schulz, et al, “Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture,” IEEE Commun. Mag., vol. 55, no. 2, pp. 70-78, Feb 2017.
[33] Fayaz, M., Yi, W., Liu, Y., & Nallanathan, A. (2020). *Transmit Power Pool Design for Grant-Free NOMA-IoT Networks via Deep Reinforcement Learning*.
[34] J. Zhang, X. Tao, H. Wu, N. Zhang, and X. Zhang, “Deep reinforcement learning for throughput improvement of the uplink grant-free NOMA system,” IEEE Internet Things J., vol. 7, no. 7, pp. 6369–6379, 2020.
[35] F. Rezaei, C. Tellambura, A. A. Tadaion and A. R. Heidarpour, "Rate Analysis of Cell-Free Massive MIMO-NOMA With Three Linear Precoders," in IEEE Transactions on Communications, vol. 68, no. 6, pp. 3480-3494, June 2020, doi: 10.1109/TCOMM.2020.2978189.
[36] Bashar, Manijeh, et al. "NOMA/OMA mode selection-based cell-free massive MIMO." *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019.
[37] Y. Li and G. A. Aruma Baduge, "NOMA-Aided Cell-Free Massive MIMO Systems," in IEEE Wireless Communications Letters, vol. 7, no. 6, pp. 950-953, Dec. 2018, doi: 10.1109/LWC.2018.2841375.
[38] F. Rezaei, A. R. Heidarpour, C. Tellambura and A. Tadaion, "Underlaid Spectrum Sharing for Cell-Free Massive MIMO-NOMA," in IEEE Communications Letters, vol. 24, no. 4, pp. 907-911, April 2020, doi: 10.1109/LCOMM.2020.2966195.
[39] M. Amjad, L. Musavian and S. Aïssa, "Link-Layer Rate of NOMA with Finite Blocklength for Low-Latency Communications," 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, London, UK, 2020, pp. 1-6, doi: 10.1109/PIMRC48278.2020.9217106.
[40] Z. Chang et al., "Energy-Efficient and Secure Resource Allocation for Multiple-Antenna NOMA With Wireless Power Transfer," in IEEE Transactions on Green Communications and Networking, vol. 2, no. 4, pp. 1059-1071, Dec. 2018, doi: 10.1109/TGCN.2018.2851603.
[41] 3rd Generation Partnership Project (3GPP), 2017, TR.38.802 “Study on New Radio Access Technology Physical Layer Aspects”, Release 14 (v14.2.0)
[42] 3rd Generation Partnership Project (3GPP), 2018, TR38.812 “Study on Non-Orthogonal Multiple Access (NOMA) for NR”, Release 16 (v16.0.0)
[43] 3rd Generation Partnership Project (3GPP), 2013, TR 36.888 “Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE”, Release 12 v(12.0.0)
[44] S. Schiessl, M. Skoglund and J. Gross, "NOMA in the Uplink: Delay Analysis With Imperfect CSI and Finite-Length Coding," in *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 3879-3893, June 2020, doi: 10.1109/TWC.2020.2979114.
[45] Z. Yang, Z. Ding, P. Fan and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems", *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244-7257, Nov. 2016.
[46] S. M. R. Islam, N. Avazov, O. A. Dobre and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges", *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721-742, 2nd Quart. 2017
[47] X. Chen, A. Benjebbour, A. Li and A. Harada, "Multi-user proportional fair scheduling for uplink non-orthogonal multiple access (NOMA)", *Proc. IEEE 79th Veh. Technol. Conf. (VTC Spring)*, pp. 1-5, May 2014.

[48] Y. Endo, Y. Kishiyama and K. Higuchi, "Uplink non-orthogonal access with MMSE-SIC in the presence of inter-cell interference", *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, pp. 261-265, Aug. 2012.
[49] N. Zhang, J. Wang, G. Kang and Y. Liu, "Uplink nonorthogonal multiple access in 5G systems", *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 458-461, Mar. 2016.
[50] R. Kassab, O. Simeone and P. Popovski, "Coexistence of URLLC and eMBB Services in the C-RAN Uplink: An Information-Theoretic Study," 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 2018, pp. 1-6, doi: 10.1109/GLOCOM.2018.8647460.

List of abbreviations

Term	Explanation	Term	Explanation
3GPP	3rd Generation Partnership Project	LORA	Long range WAN
5G-PPP	5G-public private partnership	M2M	Machine to machine
6G	6th generation	MEC	Mobile edge computing
AE	Application services	mMTC	Massive Machine Type Communications
AI/ML	Artificial Intelligence / Machine learning	MQTT	MQ Telemetry Transport
AIOTI	Alliance of IoT innovation	NFVI	Network function virtualization infrastructure
AP	Access point	NGIoT	Next generation IoT
API	Application programming interface	NGMN	Next generation mobile networks
BBU	Baseband unit	NIC	Network interface card
BS	Base station	NOMA	Non orthogonal multiple access
COAP	Constrained Application Protocol	NR	New Radio
COTS	Commercial off the shelf	OMA	Orthogonal multiple access
C-RAN	Cloud RAN	PHY	Physical layer
CSE	common services entities	PNF	Physical network function
CSI	Channel state indication	QoS	Quality of service
D2D	Device to device	RAN	Radio Access Network
DPI	Deep packet inspection	RAT	Radio access technology
DQN	Deep Q-networks	RB	Resource block
E2E	End-to-end	RRH	Remote radio head
eMBB	enhanced Multimedia Broadcast	RRM	Radio resource management
ETSI	European telecommunications standards institute	SC	Superposition coding
FDMA	Frequency division multiple access	SDR	Software defined radio

FPGA	Field programmable array	SIC	Successive interference cancellation
GPU	Graphical processing unit	SNR	Signal to noise ratio
HLA	High level architecture	SPS	Semi-persistent scheduling
HW	Hardware	TBS	Transport block size
IEEE	Institute of Electrical and Electronics Engineers	TIA	Telecommunications Industry Association
IIRA	Industrial internet reference architecture	UAV	Unmanned air vehicle
IM	Interference mitigation	UE	User equipment
IoT	Internet of Things	URLLC	Ultra low latency communications
ITU	International telecommunication union	V2X	Vehicle to everything
		VNF	Virtual network function